Recent progress with large-scale ab initio calculations: the CONQUEST code

D. R. Bowler[*1,2], **R. Choudhury**[1], **M. J. Gillan**[**1], and **T. Miyazaki**[***2]

[1] Dept. of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK
[2] National Institute for Materials Science, 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan

Received 1 October 2005

Key words List, of, comma, separated, keywords.
PACS 04A25

While the success of density functional theory (DFT) has led to its use in a wide variety of fields such as physics, chemistry, materials science and biochemistry, it has long been recognised that conventional methods are very inefficient for large complex systems, because the memory requirements scale as N^2 and the cpu requirements as N^3 (where N is the number of atoms). The principles necessary to develop methods with linear scaling of the cpu and memory requirements with system size ($\mathcal{O}(N)$ methods) have been established for more than ten years, but only recently have practical codes showing this scaling for DFT started to appear. We report recent progress in the development of the CONQUEST code, which performs $\mathcal{O}(N)$ DFT calculations on parallel computers, and has a demonstrated ability to handle systems of over 10,000 atoms. The code can be run at different levels of precision, ranging from empirical tight-binding, through *ab initio* tight-binding, to full *ab initio*, and techniques for calculating ionic forces in a consistent way at all levels of precision will be presented. Illustrations are given of practical CONQUEST calculations in the strained Ge/Si(001) system.

1 Introduction

This paper aims to summarize recent progress in techniques for performing *ab initio* calculations with computational effort scaling linearly with system size. The target systems are intended to be very large (tens of thousands or hundreds of thousands of atoms), with the implementation using methods based on density-functional theory (DFT) and pseudopotentials [1]. The CONQUEST DFT code[2, 3, 4, 5, 6, 7, 8] is designed to perform this kind of calculation, and its $\mathcal{O}(N)$ capabilities have been tested on systems of up to 16,000 atoms[7], but up to now it has been limited to rather simple systems. We report here the developments which now make it possible to do $\mathcal{O}(N)$ DFT calculations on non-trivial materials problems using the CONQUEST code.

The Car-Parrinello paper of 1985 [9] set a completely new agenda for computational condensed matter science. In that paper, the strategy based on density functional theory (DFT), pseudopotentials and plane waves was formulated in a new and powerful way that eventually led to its current widespread application across all the disciplines that are based on a molecular view of matter. Yet soon after 1985 an important limitation of the strategy had been recognised. In several papers in the early 1990's [10, 11, 12, 13], it was pointed out that the number of computer operations demanded by the Car-Parrinello strategy would increase ultimately as N^3 with the number of atoms N in the system. At the same time, it was realised that such a rapid increase must be avoidable, since the locality of quantum coherence [14, 15] (known as

[*] e-mail: david.bowler@ucl.ac.uk
[**] e-mail: m.gillan@ucl.ac.uk
[***] e-mail: miyazaki.tsuyoshi@nims.go.jp

the principle of near-sightedness[15]) should make it possible to determine the electronic ground state of a system with a number of operations that increases only linearly with N. Ideas for reformulating the strategy to achieve $\mathcal{O}(N)$ scaling were proposed at that time [11, 16, 17, 12, 13, 15, 3] , and rapidly led to practical $\mathcal{O}(N)$ methods within tight-binding schemes. The feasibility of performing $\mathcal{O}(N)$ DFT calculations on systems of many thousands of atoms was also demonstrated about 8 years ago[4]. However, the early promise of those ideas has taken a long time to be turned into practical techniques for performing $\mathcal{O}(N)$ DFT calculations on systems containing many thousands of atoms. We report here encouraging new results from the $\mathcal{O}(N)$ CONQUEST code [3, 2, 8] on systems of nearly 23,000 atoms, which, together with other recent progress, suggest that the promise is at last being realised.

We start by recalling the principles on which CONQUEST and a number of other $\mathcal{O}(N)$ codes (such as ONETEP, OpenMX and SIESTA) are based [3, 2, 8, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27]. The locality of quantum coherence is expressed by the decay of the Kohn-Sham density matrix $\rho(\mathbf{r}, \mathbf{r}') \rightarrow 0$ as $|\mathbf{r} - \mathbf{r}'| \rightarrow \infty$ [14, 15]. One way of determining the self-consistent DFT ground state is to write the total energy E_{tot} in terms of $\rho(\mathbf{r}, \mathbf{r}')$, and to minimise E_{tot} with respect to ρ, subject to the conditions that (i) ρ is Hermitian; (ii) ρ is idempotent (i.e. it is a projector); and (iii) ρ gives the correct number of electrons [3, 2]. Since E_{tot} is variational, $\mathcal{O}(N)$ operation is achieved by imposing the truncation $\rho(\mathbf{r}, \mathbf{r}') = 0$ for $|\mathbf{r} - \mathbf{r}'| > r_{\mathrm{c}}$, and the true DFT ground state is recovered as the cut-off distance r_{c} is increased.

To obtain a workable scheme, it is required that ρ be separable (i.e. that the number of its non-zero eigenvalues be finite):

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_{i\alpha, j\beta} \phi_{i\alpha}(\mathbf{r}) K_{i\alpha, j\beta} \phi_{j\beta}(\mathbf{r}') \ . \tag{1}$$

In practical codes, the $\phi_{i\alpha}(\mathbf{r})$ have been chosen to be "localised orbitals" centred on the atoms, with $\phi_{i\alpha}$ being the αth orbital on atom i. (Centring on atoms is not essential, and the orbitals can be allowed to float, see for example [28].) The localised orbitals (which in the CONQUEST scheme are referred to as "support functions") must be represented in terms of basis functions. The choice of basis function is a well-known problem in electronic structure; in this context, the two extremes are those of small basis size and simplicity (solutions of the atomic Schrödinger equation) and large basis size and systematicity (allowing an increase in accuracy). In the former group are numerical atomic orbitals[29, 30, 31, 19, 21, 32] while in the latter are spherical waves[33, 23], periodic sinc functions[25, 26], wavelets[34, 35], numerical representation on a grid[36] and blips functions[37]. Recently, there has been work showing that it is possible to define a systematically convergent set of numerical orbitals (with respect to number of orbitals per angular momentum channel)[21, 32], and that numerical orbitals can be fitted systematically to results of plane wave calculations[30]. The elements of the matrix $K_{i\alpha,j\beta}$ are then the elements of the density matrix in the (generally non-orthogonal) representation of the $\{\phi_{i\alpha}\}$. To achieve $\mathcal{O}(N)$, the $\phi_{i\alpha}(\mathbf{r})$ are required to be non-zero only within finite regions, which can be chosen as spheres of radius R_{reg}, and $K_{i\alpha,j\beta}$ is also subject to a spatial cut-off. The ground state is then found by minimising E_{tot} with respect to the orbitals $\phi_{i\alpha}(\mathbf{r})$, if the basis set allows this, and with respect to the $K_{i\alpha,j\beta}$, subject to idempotency and fixed electron number [3, 2, 8]. There are many ways of finding the ground state density matrix, $K_{i\alpha,j\beta}$, including the auxiliary density matrix method[16, 38], penalty functional methods[15, 22], the Fermi operator expansion (FOE)[39], bond-order potential (BOP)[40, 41, 20], and the constrained search formalism[42, 17, 12]. Many of these have been compared for different materials[43, 44] and are described in an extensive review article[45].

2 Overview Of Conquest Methodology

The practical implementation of any $\mathcal{O}(N)$ scheme raises two important questions initially: first, which basis set to choose for representing the localised orbitals (the most common choices were mentioned above); and second, how to find the ground state density matrix within the requirement of idempotency (again, several of the more common schemes were mentioned above). At present, in CONQUEST, two basis sets are available: the B-splines, or blips, which can be systematically converged to a given plane wave result; and pseudo-atomic orbitals, which are economical and very often give accurate results. We envisage that we will implement other options as required. The main emphasis in the present paper is on pseudo-atomic orbitals.

As mentioned above, there are various techniques in use for finding the ground-state density matrix. The present implementation in CONQUEST[46] is a combination of the techniques of Li, Nunes and Vanderbilt (LNV)[16] and Palser and Manolopoulos [47], both of which are closely related to McWeeny's 'purification' scheme [48]. In the LNV technique, the density matrix K is represented in terms of an 'auxiliary' density matrix L as:

$$K = 3LSL - 2LSLSL \,, \tag{2}$$

where S is the overlap matrix for the support functions or localised orbitals: $S_{\lambda\mu} = \langle \phi_\lambda | \phi_\mu \rangle$. This scheme enforces 'weak' idempotency [17] (meaning that all eigenvalues are in the interval $[0, 1]$) rather than strict idempotency. If the total energy is minimised with respect to L (writing $E_{\mathrm{GS}} = 2\mathrm{Tr}[KS]$ for the band energy), this scheme automatically drives K towards idempotency. In order to ensure $\mathcal{O}(N)$ scaling, a spatial cut-off is imposed on the L-matrix, so that $L_{\lambda\mu} = 0$ when the distance between the centres of the support-functions ϕ_λ and ϕ_μ exceeds a chosen cut-off R_L. Other methods for finding the ground-state density matrix could be implemented with little effort.

As well as operating in the $\mathcal{O}(N)$ mode, CONQUEST can find the ground state directly by diagonalisation, using the SCALAPACK package, which allows efficient parallelisation of the diagonalisation. Since it scales as $\mathcal{O}(N^3)$, this will only be appr記rpriate for relatively small systems, but it provides an important tool both for testing the outer parts of the ground-state search (described below) and for exploring the convergence of the $\mathcal{O}(N)$ algorithm with the cut-off on the L-matrix.

The search for the ground-state is organised into three loops. In the innermost loop, the support functions and electron density are fixed and the ground-state density matrix is found, either by varying L or by diagonalisation. In the middle loop, self-consistency is achieved by systematically reducing the electron-density residual, i.e. the difference between the input and output density in a given self-consistency cycle [49]. In the outer loop, the energy is minimised with respect to the support functions, ϕ_λ. This organisation corresponds to a hierarchy of approximations: when the inner loop alone is used, we get the scheme known as non-self-consistent *ab initio* tight binding (NSC-AITB), which is a form of the Harris-Foulkes approximation [50, 51, 29, 52]; when the inner two loops are used, we get self-consistent *ab initio* tight binding (SC-AITB); finally, if all loops are used, we have full *ab initio*. In this last case, we recover the exact DFT ground state as the region radius R_{reg} and the L-matrix cut-off R_L are increased. For non-metallic systems, the evidence so far is that accurate approximations to the ground state are obtained with quite modest values of the cut-offs [3, 19].

For calculations at the level of full *ab initio* accuracy, the convergence of the outer loop (optimising the support functions with respect to their basis functions) is well-conditioned provided appropriate preconditioning measures are taken; these have been discussed both for blips in the context of CONQUEST[53, 54] and for psinc functions in the context of ONETEP[26]. We note that CONQUEST can be run in a mode analagous to SIESTA, where pseudo-atomic orbitals are used and no optimisation is performed; in this case, the outer loop is not performed.

We have recently found that the self-consistency search (the middle loop described above) can be accelerated by use of the Kerker preconditioning. This idea, which is well-known in the plane-wave community, removes long wavelength changes in the charge density during mixing. It is applied in reciprocal space, as a prefactor:

$$f(q) = \frac{q^2}{q^2 + q_0^2} \tag{3}$$

Then the charge is mixed using a Pulay or Broyden (or related) scheme[49] with the prefactor applied to the residual or output charge after transformation to reciprocal space. The mixing includes a parameter, A, which determines how aggressive the mixing is (with the input charge density for iteration $n+1$ given by $\rho^{n+1}_{in} = \rho^{n}_{in} + Af(q)R_n$, with R_n the residual from iteration n). The efficacy of this preconditioning is explored, both for exact diagonalisation and linear scaling, in Section 4.1.

While performing the search for self-consistency, we must monitor the residual. We define the following dimensionless parameter which is used to monitor the search:

$$d = \frac{\langle |R(\mathbf{r})|^2 \rangle^{1/2}}{\bar{\rho}}, \qquad (4)$$

$$\langle |R(\mathbf{r})|^2 \rangle = \frac{1}{V} \int \mathrm{d}\mathbf{r}\, |R(\mathbf{r})|^2 , \qquad (5)$$

where V is the simulation cell volume and we use the usual definition of residual, $R(\mathbf{r}) = \rho_{\mathrm{out}}(\mathbf{r}) - \rho_{\mathrm{in}}(\mathbf{r})$, the difference between the output and input charge densities. The quantity d is then the RMS value of $R(\mathbf{r})$ normalised by dividing by the *average* charge density in the system, $\bar{\rho}$. Note that, for systems containing large amounts of vacuum, the criterion for convergence will need to be altered when compared to bulk-like environments. This criterion may be coupled with a monitor on the largest value of residual on an individual grid point $\mathbf{r}_l$, $R_{\mathrm{max}} = \mathrm{max}_l\, |R(\mathbf{r}_l)|$

The scheme we have outlined is closely related to the methods used in SIESTA [18, 19], OpenMX[21] and ONETEP[27]. The main differences are: (i) the basis sets chosen (SIESTA uses fixed PAOs, while OpenMX uses optimized orbitals and ONETEP psinc functions); (ii) the method of finding the ground state density matrix (Siesta uses the constrained search technique[17, 12, 42], OpenMX the divide-and-conquer[55] or BOP[20] and ONETEP either penalty functional[15, 33] or LNV[16]); (iii) the technique of 'neutral-atom potentials' [18, 19], used by SIESTA and OpenMX, which allows calculation of matrix elements to be performed very efficiently for localised, atomic-like basis sets.

The principle of near-sightedness (i.e. spatial locality of electronic structure), means that sparse matrix multiplications and other operations are restricted to a finite area of space, leading to a natural route to parallel decomposition of the problem. CONQUEST was written from the outset as parallel code, and a large part of the development effort has been concerned with techniques for achieving good parallel scaling. The parallelisation techniques have been described in detail elsewhere [4, 7, 8], so we give only a brief summary. There are three main types of operation that must be carefully distributed across processors:

- the storage and manipulation of localised orbitals, e.g. the calculation of $\phi_\lambda(\mathbf{r})$ on the integration grid starting from blip- or PAO-coefficients, and the calculation of the derivatives of E_{tot} with respect to these coefficients, which are needed for the ground-state search;
- the storage and manipulation of elements of the various matrices (H, S, K, L, etc...);
- the calculation of matrix elements by summation over domains of points on the integration grid, or by analytic operations (for certain integrals involving PAOs and blips).

Efficient parallelisation of these operations, and the elimination of unnecessary communication between processors, depend heavily on the organisation of both atoms and grid points into small compact sets, which are assigned to processors [7]. When the code runs in $\mathcal{O}(N)$ mode, matrix multiplication takes a large part of the computer effort, and we have developed parallel multiplication techniques [7] that exploit the specific patterns of sparsity on which $\mathcal{O}(N)$ operation depends.

3 Calculation of Ionic Forces

In order to perform structural relaxation or molecular dynamics of materials with an electronic structure technique, the algorithms for calculating the forces $\mathbf{F}_i$ on the ions must be the exact derivatives of the total ground state energy, E_{GS}, with respect to the positions, $\mathbf{r}_i$, such that $\mathbf{F}_i = -\nabla_i E_{\mathrm{GS}}$. One of the advantages of DFT, within the pseudopotential approximation, is that it is easy, in principle, to achieve this relationship between the forces and the energy. Since the CONQUEST formalism allows the calculation of the total energy at different levels of accuracy, some care is needed in the formulation of the forces to develop a scheme that works at all levels of this hierarchy. It is also important to ensure that it works equally well (and accurately) for both the diagonalisation and $\mathcal{O}(N)$ modes of operation implemented in CONQUEST. We have recently described these algorithms in detail[56], but we summarise them below for convenience.

We recall the Harris-Foulkes expression[50, 51] for the total energy, which is often applied when self-consistency is not sought, but which at self-consistency is identical to the standard Kohn-Sham expression for total energy. The expression is:

$$E_{\mathrm{GS}} = E_{\mathrm{BS}} + \Delta E_{\mathrm{Har}} + \Delta E_{\mathrm{xc}} + E_{\mathrm{C}}, \tag{6}$$

with E_{C} the Coulomb energy between the ionic cores, and the band-structure energy, the double-counting Hartree and exchange-correlation energies defined as:

$$\begin{aligned} E_{\mathrm{BS}} &= 2\sum_n f_n \epsilon_n && (7)\\ &= 2\mathrm{Tr}[KH] && (8)\\ \Delta E_{\mathrm{Har}} &= -\frac{1}{2}\int d\mathbf{r}\, n^{\mathrm{in}}(\mathbf{r}) V^{\mathrm{in}}_{\mathrm{Har}}(\mathbf{r}) \\ \Delta E_{\mathrm{xc}} &= \int d\mathbf{r}\, n^{\mathrm{in}}(\mathbf{r}) \left(\epsilon_{\mathrm{xc}}(n^{\mathrm{in}}(\mathbf{r})) - \mu_{\mathrm{xc}}(n^{\mathrm{in}}(\mathbf{r}))\right) \,. && (9) \end{aligned}$$

Here, $n^{\mathrm{in}}(\mathbf{r})$ is the *input* charge density used (normally a superposition of atomic charge densities if a non-self-consistent scheme is used, or the self-consistent charge density if self-consistency is used). This expression is very useful when comparing forces at different levels of approximation.

At the empirical TB level, the ionic force is a sum of the band-structure part $\mathbf{F}_i^{\mathrm{BS}}$ and the pair-potential part $\mathbf{F}_i^{\mathrm{pair}}$, the former being given by [29]:

$$\mathbf{F}_i^{\mathrm{BS}} = -2\mathrm{Tr}\left[K\nabla_i H - J\nabla_i S\right] , \tag{10}$$

where K and J are the density matrix and energy matrix respectively [29]. It is readily shown that in the $\mathcal{O}(N)$ scheme of LNV, and in some other $\mathcal{O}(N)$ schemes, the same formula for $\mathbf{F}_i^{\mathrm{BS}}$ is the exact derivative of the $\mathcal{O}(N)$ total energy. In the LNV scheme, K is given by eqn (2), and J by:

$$J = -3LHL + 2LSLHL + 2LHLSL \,. \tag{11}$$

In NSC-AITB (Harris-Foulkes), the forces can be written in two equivalent ways. The way that corresponds most closely to empirical TB is:

$$\mathbf{F}_i = \mathbf{F}_i^{\mathrm{BS}} + \mathbf{F}_i^{\Delta\mathrm{Har}} + \mathbf{F}_i^{\Delta\mathrm{xc}} + \mathbf{F}_i^{\mathrm{ion}} \;, \tag{12}$$

where $\mathbf{F}_i^{\mathrm{BS}}$ is given by exactly the same formula as in empirical TB. The contributions $\mathbf{F}_i^{\Delta\mathrm{Har}}$ and $\mathbf{F}_i^{\Delta\mathrm{xc}}$, which arise from the double-counting Hartree and exchange-correlation parts of the NSC-AITB total energy, have been discussed elsewhere [29]. The final term $\mathbf{F}_i^{\mathrm{ion}}$ come from the ion-ion Coulomb energy. This way of writing $\mathbf{F}_i$ expresses the well-known relationship between NSC-AITB and empirical TB that in the latter the pair term represents the sum of the three contributions $\Delta\mathrm{Har} + \Delta\mathrm{xc} + \mathrm{ion} - \mathrm{ion}$. The alternative, and exactly equivalent, way of writing $\mathbf{F}_i$ in NSC-AITB is:

$$\mathbf{F}_i = \mathbf{F}_i^{\mathrm{ps}} + \mathbf{F}_i^{\mathrm{Pulay}} + \mathbf{F}_i^{\mathrm{NSC}} + \mathbf{F}_i^{\mathrm{ion}} \; . \tag{13}$$

Here, $\mathbf{F}_i^{\mathrm{ps}}$ is the "Hellmann-Feynman" force exerted by the valence electrons on the ion cores; $\mathbf{F}_i^{\mathrm{Pulay}}$ is the Pulay force that arises in any method where the basis set depends on ionic positions; $\mathbf{F}_i^{\mathrm{NSC}}$ is a force contribution associated with non-self-consistency, and is expressed in terms of the difference between output and input electron densities; $\mathbf{F}_i^{\mathrm{ion}}$, as before, is the ion-ion Coulomb force. Exactly the same formulas represent the exact derivative of E_{tot} in both diagonalisation and $\mathcal{O}(N)$ modes.

In both SC-AITB and full AI, the force formula is:

$$\mathbf{F}_i = \mathbf{F}_i^{\mathrm{ps}} + \mathbf{F}_i^{\mathrm{Pulay}} + \mathbf{F}_i^{\mathrm{ion}} \; , \tag{14}$$

which differs from the second version of the NSC-AITB formula eqn (13) only by the absence of the non-self-consistent contribution $\mathbf{F}_i^{\mathrm{NSC}}$, as expected.

The above hierarchy of force formulas has been implemented in CONQUEST, and extensive tests have ensured that the total energy and the forces are exactly consistent within rounding-error precision[56].

4 Illustrative Results

We have already demonstrated the ability of Conquest to address non-trivial systems by relaxing the Si(001) surface with a variety of basis sets, and comparing the results to geometries obtained with both Siesta and VASP[56]. In this section, we will present further results which show that CONQUEST is now ready for application to real-world scientific problems.

We start with the problem of self-consistency, and demonstrate both that the Kerker preconditioning technique described in Section 2 above is extremely useful and that it does not degrade the linear scaling stability. We then present selected data from a study of the three-dimensional "huts" which evolve to relieve strain during heteroepitaxial growth of Ge on Si(001).

4.1 Self-consistency

The search for self-consistency between the charge density and potential can still sometimes be problematic in standard DFT codes: even with the most sophisticated, convergence to the ground state is not completely guaranteed. In this section, we present the results of tests on three different systems, demonstrating the effects of varying the mixing parameter, A, and the preconditioning wavevector, q_0 in the Kerker preconditioning, as discussed in Section 2. The results will be both for exact diagonalisation and linear scaling ground state search methodologies, though we reserve the linear scaling results for the most challenging problem, presented last. We present the raw data (change of residual with iteration) for selected cases at the end of the section, along with data on the effect of region radius on convergence rate.

The parameters used in modelling the systems are as follows: CONQUEST was operating with a PAO basis set at the single-zeta level (with four orbitals per atom) following the generation scheme in SIESTA, with cutoffs of 4.88 and 6.12 bohr on the s- and p-orbitals respectively (corresponding to a shift of 250 meV on-site). There was no outer loop (as discussed in Sec. 2 above). The criterion applied for reaching self-consistency was d=0.01 for bulk silicon (where d is defined in Eq. (4) above). This value of d is perfectly adequate for total energy calculations though may be a little loose for accurate molecular dynamics; for this purpose it is ideal as it is the early stages of the self-consistency search which are non-linear. For both the silicon surface and the cluster, the *average* electron density in the system is smaller owing to the vacuum present; in both cases we used larger values of d so that the effective convergence was the same (for Si(001) we set d=0.015 and for the silicon cluster d=0.039). In all cases, the starting charge density was a linear superposition of atomic charge densities generated from the PAO basis functions.

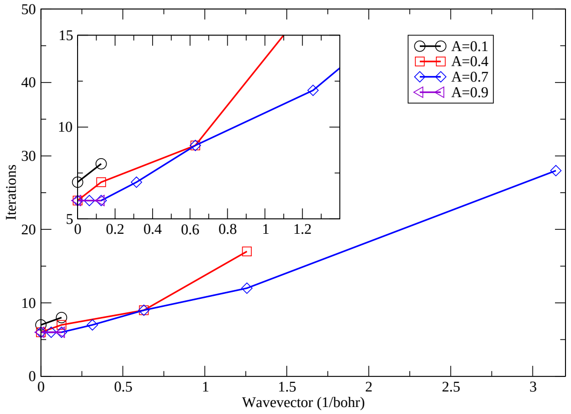

Fig. 1 Convergence to self-consistency with Kerker preconditioning wavevector and mixing parameter, A, for 512 atom cell of bulk silicon. Any value of 50 iterations indicates a lack of convergence for that parameter.

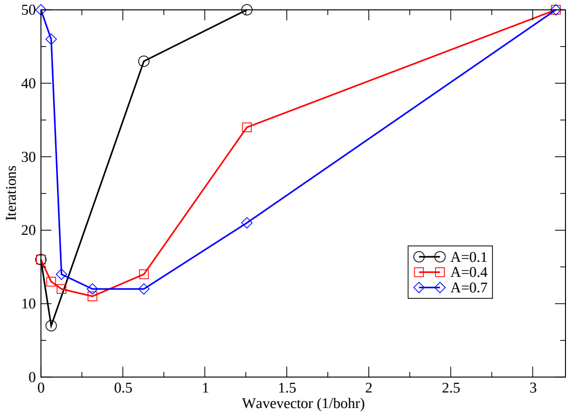

Fig. 2 Convergence to self-consistency with Kerker preconditioning wavevector and mixing parameter, A, for 192 atom cell of Si(001). Any value of 50 iterations indicates a lack of convergence for that parameter.

We begin with the simplest possible system: bulk silicon. The results for a unit cell containing 512 atoms are given in Fig. 1, purely for diagonalisation at the gamma point. As might be expected for such a simple system, there is no problem reaching the ground state, and in fact the Kerker preconditioning is ineffective: at best it has no effect, and it often slows down the iterations.

We now move from a three-dimensional system to the Si(001) surface. The results for a slab containing 192 atoms (the unit cell was one dimer row long, eight dimers long, and 12 layers deep with a surface on both sides) are given in Fig. 2. The Kerker preconditioning is vital here, with the value of q_0 determining stability during the search.

Finally we turn to an amorphous cluster of 343 silicon atoms, whose structure was optimised by a minima-hopping procedure[57]. A projection of the structure of the cluster is shown in Fig. 3, along with the convergence results. The results for exact diagonalisation have been reproduced with linear scaling to test the effect of Kerker mixing on the stability of the algorithm (shown with downward triangles). As is clear from the plot, the use of linear scaling for this cluster is not affected at all by the choice of self-consistency algorithm, and convergence *to self-consistency* is equally fast with either method. We note that the cluster is metallic (or at least that there is partial occupation of a number of levels near the Fermi level),

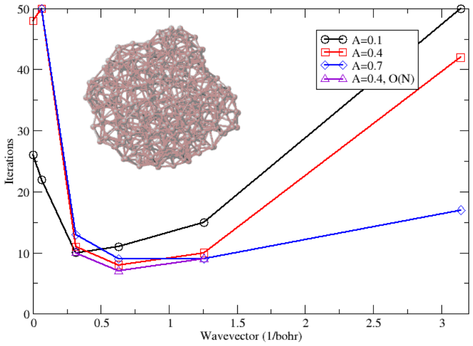

Fig. 3 Convergence to self-consistency with Kerker preconditioning wavevector and mixing parameter, A, for amorphous, metallic 343 atom cluster of silicon. Any value of 50 iterations indicates a lack of convergence for that parameter.

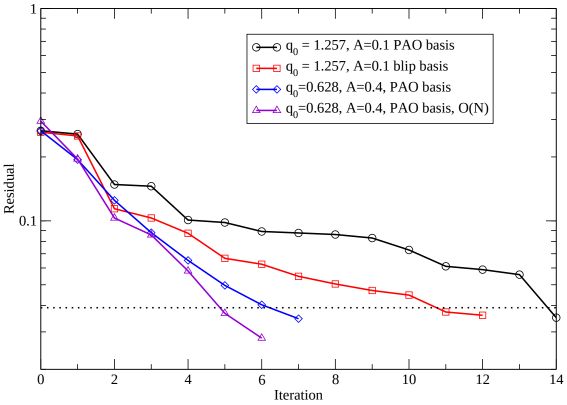

Fig. 4 Residual, defined in Eq. (4), during self-consistency search. The system considered is the amorphous cluster shown in Fig. 3, and different conditions are considered for the simulation as discussed in the main text.

leading to rather slow convergence of the density matrix minimisation (though we note that the cutoff on the density matrix was 7 Å, which is not large enough to provide convergence to the diagonalisation result). Nevertheless, the self-consistent ground state is achieved without significant difficulty.

We note that the optimal values of q_0 found are in good agreement with those previously published, though there is little quantitative data available beyond Ref. [58]. In that publication, it was stated that values of q_0 in the range 0.5—2.0 Å^{-1} yielded little change in the convergence rate. The factor to convert between bohr^{-1} and Å^{-1} which is required in order to compare this range to the data in Figs. 1–3 is 1.8897 (simply one bohr^{-1}). Then we see that, excluding bulk silicon, the acceptable range of 0.20–0.65 bohr^{-1} for Si(001) and 0.30–1.25 bohr^{-1} for the cluster convert to 0.38–1.23 Å^{-1} and 0.57–2.36 Å^{-1}, in very good agreement. We also note that the problem of slow convergence for overly small mixing parameter A noted[58] is seen in Fig. 2.

The raw data, that is value of d per iteration, is shown in Fig. 4 for the 343 atom cluster. For q_0=0.2 and A=0.1, we show convergence for the single zeta PAO basis (circles) and an *unconverged* blip basis

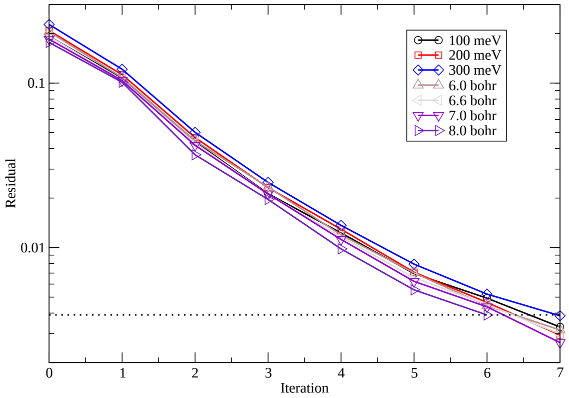

Fig. 5 Residual, defined in Eq. (4), during self-consistency search for different region radii. The system considered is the amorphous cluster shown in Fig. 3, with the convergence used for that figure, and Kerker values A=0.4, q_0=0.628 bohr^{-1}. Details of basis sets are given in the text.

Table 1 Surface energy for Ge(105) calculated using exact diagonalisation (labelled diag) and linear scaling with different density matrix cutoff distances.

R_L(bohr)	$\mathrm{E}_{\mathrm{tot}}$ (Ha)	$\mathrm{F}_{\mathrm{max}}$ (Ha/bohr)	$\mathrm{E}_{\mathrm{surf}}$(eV/A^2)
15.4	-484.7635	0.0027	0.0801
20.4	-485.0438	0.0014	0.0753
25.4	-485.1420	0.0007	0.0752
30.4	-485.1811	0.0004	0.0755
diag	-485.2048	0.0001	0.0765

(squares). This shows that the convergence rate is only mildly dependent on choice of basis, or indeed on the convergence of the outer loop (since the blips chosen were far from optimal). Similarly, for q_0=0.1 and A=0.4, we show convergence for both exact diagonalisation (diamonds) and linear scaling (triangles) solvers . Again, the rate of convergence is only mildly changed by this change of methodology. These results are illustrative of the general behaviour of the minimisation: the self-consistency scheme described is unaffected by the details of the inner or outer loops.

Finally, we present further a further plot of the residual per iteration for the 343 atom Si cluster for different values of the region radius in Fig. 5. For this test, we have again used single-zeta PAOs, with different techniques for generation. We used both the SIESTA energy shift (yielding radii of 5.67/7.11 bohr for s/p with 100 meV, 5.26/6.43 bohr for 200 meV and 4.88/6.12 bohr for 300 meV) and fixed radii for both s and p channels (of 6.0 bohr, 6.6 bohr, 7.0 bohr and 8.0 bohr). We see that there is only a very small dependence of convergence on region radius.

4.2 Hut clusters of Ge on Si(001)

We are currently investigating the stability of three-dimensional islands of Ge on Si(001) during strained heteroepitaxial growth[59], following our earlier studies of strained layer growth using conventional DFT[60] and linear scaling tight binding[61]. After approximately three monolayers of Ge has been deposited (though this varies with the system and growth conditions), small mounds with well-defined facets begin to appear. These have facets made of the Ge(105) surface, and allow significant strain relaxation. A first step in understanding these huts is to characterise the Ge(105) surface[62], which we have done with CONQUEST, in particular concentrating on the density matrix range required for accurate calculations. We present surface energy and maximum forces calculated using a minimal basis at the NSC-AITB level for a

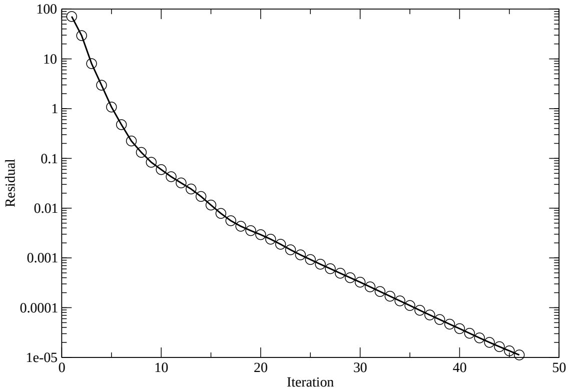

Fig. 6 Convergence during minimisation of energy with respect to density matrix elements (innermost loop of ground state search) for 23,000 atom Ge hut cluster on Si(001) system.

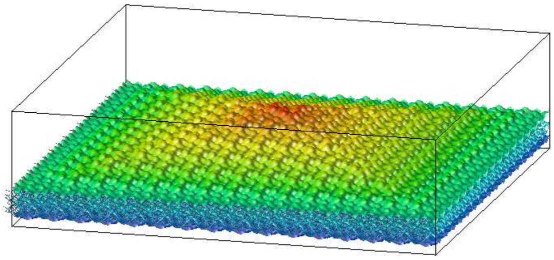

Fig. 7 Charge density from 23,000 atom Ge hut cluster on Si(001) system. The density was sampled at one grid point in four in each direction to create a manageable data set. Colour indicates height; isosurface is equivalent to ∼0.3 electrons per cubic Ångstrom.

Ge(105) surface in Table 4.2. (We note that the structure modelled here is a previously unreported insulating structure which will be presented in more detail elsewhere.) The surface energy was calculated using systems with eight layers and ten layers (100 and 124 atoms respectively) The system was relaxed using exact diagonalisation (with a $4 \times 5 \times 1$ Monkhorst-Pack mesh) and this structure was used in the linear scaling calculations. We see that the maximum force is converged at a density matrix range of 20 or 25 bohr, but that the surface energy is more slowly convergent. Nevertheless, this is a strong indication that reliable results can be obtained using linear scaling[1].

We have considered a number of different hut cluster systems, with different sizes and spacings on the substrate. The largest system which we have prepared so far contains approximately 23,000 atoms (including substrate, wetting layer of Ge and the hut cluster). We show the residual of the density matrix against iteration number in Fig. 6, which demonstrates both that convergence to the ground state can be easily achieved for a system this large, and that the number of iterations needed to reach the ground state is independent of system size (which is an important consideration for this type of code).

Finally, we present a preliminary charge density for this system in Fig. 7, where the colour indicates height in the unit cell. The isosurface was formed by sampling one grid point in four in each direction, to make the data set manageable. The charge density shows both the scale of the simulation and the dimers

[1] In fact, on converging the $\mathcal{O}(N)$ results it was found that the exact diagonalisation results were insufficiently converged with respect to the k-point mesh, and the mesh had to be increased.

on the facets of the hut, as well as the characteristic buckling associated both with the Si(001) surface and the Ge(105) faces of the hut cluster.

5 Discussion And Conclusions

We have presented an overview of the methodology used in the CONQUEST code, and shown that realistic calculations on complex, scientifically interesting systems are now possible. In particular, we have outlined the arrangement of the code to allow calculations at different levels of precision (from minimal-basis AITB through to full *ab initio*) and presented details of how forces can be calculated consistently at all levels of precision. We have shown details of convergence to self-consistency for various different systems, and outlined early results from our study of three-dimensional "hut" clusters resulting from heteroepitaxial growth of Ge on Si(001).

We want to note that there are other $\mathcal{O}(N)$ techniques within the DFT arena, notably Siesta[19], OpenMX[21, 32] and ONETEP[27]. However, the ideas of linear scaling are not limited to DFT, and have been used in $\mathcal{O}(N)$ Hartree-Fock methods[63, 64, 65]. Very recently, some of these ideas have been applied to Quantum Monte Carlo techniques[66], and it has been shown that localised orbitals with blip-function basis sets are capable of giving a major speed-up to Quantum Monte Carlo calculations[67]. There are clear signs that $\mathcal{O}(N)$ methods are realising their early promise.

Acknowledgements The CONQUEST project is partially supported by ACT-JST. DRB is supported by a Royal Society University Research Fellowship, and RC by an EPSRC studentship. This study was performed through Special Coordination Funds for Promoting Science and Technology from the MEXT, Japan.

References

[1] R. M. Martin. Electronic Structure (Cambridge, 2004).
[2] E. Hernandez and M. J. Gillan. Phys. Rev. B **51**, 10157 (1995).
[3] E. Hernández, M. J. Gillan, and C. M. Goringe. Phys. Rev. B **53**, 7147 (1996).
[4] C. M. Goringe, E. Hernández, M. J. Gillan, et al. Comput. Phys. Commun. **102**, 1 (1997).
[5] D. R. Bowler and M. J. Gillan. Comput. Phys. Commun. **120**, 95 (1999).
[6] D. R. Bowler, I. J. Bush, and M. J. Gillan. Int. J. Quantum Chem. **77**, 831 (2000).
[7] D. R. Bowler, T. Miyazaki, and M. J. Gillan. Computer Physics Communications **137**, 255 (2001).
[8] D.R.Bowler, T.Miyazaki, and M.J.Gillan. J. Phys.:Condens. Matter **14**, 2781 (2002).
[9] R. Car and M. Parrinello. Phys. Rev. Lett. **55**, 2471 (1985).
[10] S. Baroni and P. Giannozzi. Europhys. Lett. **17**, 547 (1991).
[11] G. Galli and M. Parrinello. Phys. Rev. Lett. **69**, 3547 (1992).
[12] J. Kim, F. Mauri, and G. Galli. Phys. Rev. B **52**, 1640 (1995).
[13] P. Ordejón, D. Drabold, R. Martin, et al. Phys. Rev. B **51**, 1456 (1995).
[14] W. Kohn. Phys. Rev. B. **115**, 809 (1959).
[15] W. Kohn. Phys. Rev. Lett. **76**, 3168 (1996).
[16] X.-P. Li, R. W. Nunes, and D. Vanderbilt. Phys. Rev. B **47**, 10891 (1993).
[17] F. Mauri, G. Galli, and R. Car. Phys. Rev. B **47**, 9973 (1993).
[18] P. Ordejon, E. Artacho, and J. M. Soler. Phys. Rev. B **53**, R10441 (1996).
[19] J. M. Soler, E. Artacho, J. D. Gale, et al. J. Phys.:Condens. Matter **14**, 2745 (2002).
[20] T. Ozaki and K. Terakura. Phys. Rev. B **64**, 195126 (2001).
[21] T. Ozaki. Phys. Rev. B. **67**, 155108 (2003).
[22] P. D. Haynes and M. C. Payne. Phys. Rev. B **59**, 12173 (1999).
[23] C. K. Gan, P. D. Haynes, and M. Payne. Phys. Rev. B **63**, 205109 (2001).
[24] C.-K. Skylaris, A. M. Mostofi, P. D. Haynes, et al. Phys. Rev. B. **66**, 035119 (2002).
[25] A. A. Mostofi, C.-K. Skylaris, P. D. Haynes, et al. Comput. Phys. Commun. **147**, 788 (2002).
[26] A. A. Mostofi, P. D. Haynes, C.-K. Skylaris, et al. J. Chem. Phys. **119**, 8842 (2003).
[27] C.-K. Skylaris, P. D. Haynes, A. M. Mostofi, et al. J. Chem. Phys. **122**, 084119 (2005).
[28] J. Fattebert and F. Gygi. Comp. Phys. Comm **162**, 24 (2004).
[29] O. F. Sankey and D. J. Niklewski. Phys. Rev. B **40**, 3979 (1989).

[30] S. D. Kenny, A. P. Horsfield, and H. Fujitani. Phys. Rev. B **62**, 4899 (2000).
[31] J. Junquera, O. Paz, D. Sanchez-Portal, et al. Phys. Rev. B **64**, 235111 (2001).
[32] T. Ozaki and H. Kino. Phys. Rev. B. **69**, 195113 (2004).
[33] P. D. Haynes and M. C. Payne. Comput. Phys. Commun. **102**, 17 (1999).
[34] T. A. Arias. Rev. Mod. Phys. **71**, 267 (1999).
[35] S. Goedecker, T. Deutsch, X. Gonze, et al. BigDFT collaboration. For details, see `http://www-drfmc.cea.fr/sp2m/L_Sim/BigDFT/index.html`.
[36] J. L. Fattebert and J. Bernholc. Phys. Rev. B **62**, 1713 (2000).
[37] E.Hernández, M.J.Gillan, and C.M.Goringe. Phys. Rev. B **55**, 13485 (1997).
[38] R. Nunes and D. Vanderbilt. Phys. Rev. B **50**, 17611 (1994).
[39] S. Goedecker and L. Colombo. Phys. Rev. Lett. **73**, 122 (1994).
[40] D. G. Pettifor. Phys. Rev. Lett. **63**, 2480 (1989).
[41] M. Aoki. Phys. Rev. Lett. **71**, 3842 (1993).
[42] P. Ordejón, D. Drabold, M. Grumbach, et al. Phys. Rev. B **48**, 14646 (1993).
[43] D. R. Bowler, M. Aoki, C. M. Goringe, et al. Modell. Simul. Mater. Sci. Eng. **5**, 199 (1997).
[44] A. D. Daniels, J. M. Millam, and G. E. Scuseria. J. Chem. Phys. **107**, 425 (1997).
[45] S. Goedecker. Rev. Mod. Phys. **71**, 1085 (1999).
[46] D. Bowler and M. Gillan. Comp. Phys. Comm. **120**, 95 (1999).
[47] A. H. Palser and D. E. Manolopoulos. Phys. Rev. B **58**, 12704 (1998).
[48] R. McWeeny. Rev. Mod. Phys. **32**, 335 (1960).
[49] D. D. Johnson. Phys. Rev. B **38**, 12807 (1988).
[50] J. Harris. Phys. Rev. B **31**, 1770 (1985).
[51] W. Foulkes and R. Haydock. Phys. Rev. B **39**, 12520 (1989).
[52] A. P. Horsfield and A. M. Bratkovsky. J. Phys.: Condens. Matter **12**, R1 (2000).
[53] D. R. Bowler and M. J. Gillan. Comp. Phys. Commun. **112**, 103 (1998).
[54] M.J.Gillan, D.R.Bowler, C.M.Goringe, et al. In: The Physics of Complex Liquids, edited by F.Yonezawa, K.Tsuji, K.Kaji, et al. (World Scientific, 1998).
[55] W. Yang. Phys. Rev. Lett **66**, 1438 (1991).
[56] T. Miyazaki, D. Bowler, R. Choudhury, et al. J. Chem. Phys. **121**, 6186 (2004).
[57] S. Goedecker. J. Chem. Phys. **120**, 9911 (2004).
[58] G.Kresse and J.Furthmüller. Comp. Mat. Sci. **6**, 15 (1996).
[59] J. Stangl, V. Holý, and G. Bauer. Rev. Mod. Phys. **76**, 725 (2004).
[60] J. Oviedo, D. R. Bowler, and M. J. Gillan. Surf. Sci. **515**, 483 (2002).
[61] K. Li, D. R. Bowler, and M. J. Gillan. Surf. Sci. **526**, 356 (2003).
[62] Y. Fujikawa, K. Akiyama, T. Nagao, et al. Phys. Rev. Lett. **88**, 176101 (2002).
[63] M. Challacombe. J. Chem. Phys. **110**, 2332 (1999).
[64] P. Ayala and G. Scuseria. J. Chem. Phys. **110**, 3660 (1999).
[65] M. S. Lee, P. E. Maslen, and M. Head-Gordon. J. Chem. Phys. **112**, 3592 (2000).
[66] A. J. Williamson, R. Q. Hood, and J. C. Grossman. Phys. Rev. Lett. **87**, 246406 (2001).
[67] D. Alfé and M. Gillan. J. Phys. Cond. Matt. **16**, L305 (2004).